Ex. 1

1.1 $Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_p x^p + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2)$

Dataset: $\{x_i, y_i\}_{i=1}^N$

(a) $\varepsilon = Y - (\theta_0 + \theta_1 x + \dots + \theta_p x^p)$

$$p(\varepsilon) = \prod_{n=1}^{N} p(\varepsilon_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \prod_{i=1}^{N} \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

$$p(Y \mid \theta, \varepsilon) = \mathcal{N}(Y; \theta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \cdot \prod_{i=1}^{N} \exp\left(-\frac{(Y_i - (\theta_0 + \dots + \theta_p x_i^p)^2}{2\sigma^2}\right)$$

$$\rightarrow \mathcal{L}\{p(Y \mid \theta)\} = -N \log(\sqrt{2\pi}\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - (\theta_0 + \dots + \theta_p x_i^p)^2$$

$$= -N \log(\sqrt{2\pi}\sigma^2) - \frac{1}{2\sigma^2} \| \underline{Y} - \underline{\underline{X}}\,\underline{\theta} \|^2$$

where $\underline{Y} \in \mathbb{R}^N$, $\underline{\underline{X}} \in \mathbb{R}^{N \times P}$, $\underline{\theta} \in \mathbb{R}^P$

(b) Maximizing $\mathcal{L}$ in this case is equal to find the least squares solution, since the noise is gaussian

Hence $\frac{d}{d\theta} \| Y - X\theta \|^2 \stackrel{!}{=} 0 \rightarrow \frac{d}{d\theta} (Y - X\theta)^T (Y - X\theta) \stackrel{!}{=} 0$

$$\rightarrow \frac{d}{d\theta} (Y^T Y - 2 Y^T X \theta + \theta^T X^T X \theta) \stackrel{!}{=} 0 \rightarrow$$

$$\rightarrow -2 Y X^T + 2 X^T X \theta = 0 \rightarrow \boxed{\hat{\theta} = (X^T X)^{-1} X^T Y}$$

(c) $Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \varepsilon$, $\theta_0 = 0.3$, $\theta_1 = -0.1$, $\theta_2 = 0.5$
$\sigma^2 = 0.001$

$x = [-0.5 : 0.2 : 0.1]$

→ $Y = [0.4716, 0.4694, 0.3624, 0.3849, 0.2789, 0.2737, 0.2652, 0.2785]$

(d) for $P=2$: $\hat{\theta} = [0.2939, -0.1009, 0.5783]$ $L = -12,93$

for $P=1$: $\hat{\theta} = [0,3103, -0.2743]$ $L = -10,12$

for $P=7$: $\hat{\theta} = [0.3, 0.2, -1.5, -38.7, -19.6, 752.8, 2312]$
$L_7 = -7,3515$

We can notice that with too many parameters ($P=7$) the weights get very big and we completely overfit. Also it is possible that the likelihood has max-value for $P=7$.

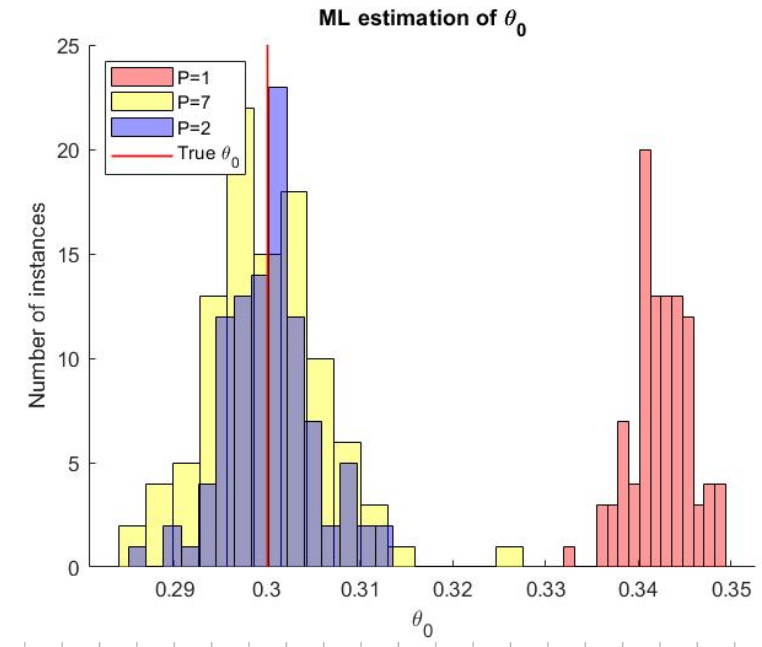
(e) For $P=1$: -376.6137 using the parameter calculated in (d)

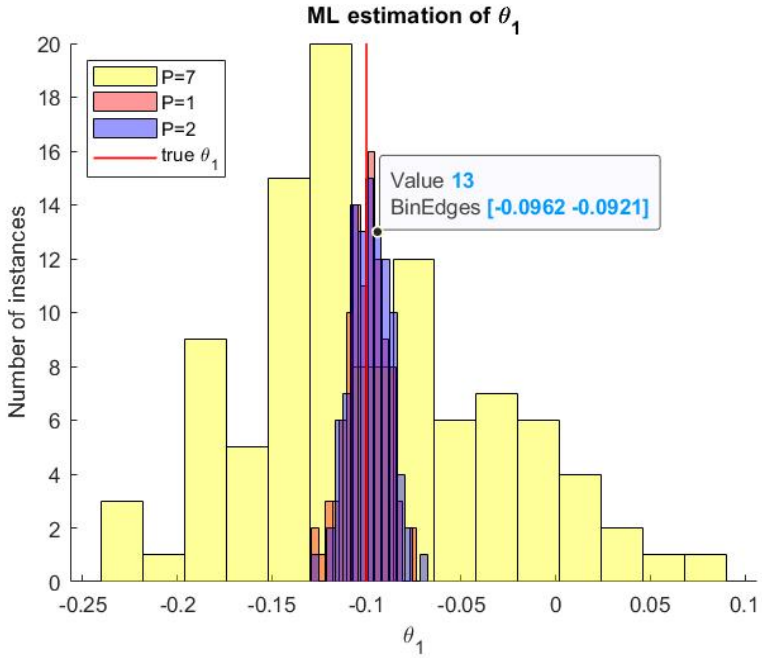
$P=2$: -148.5543

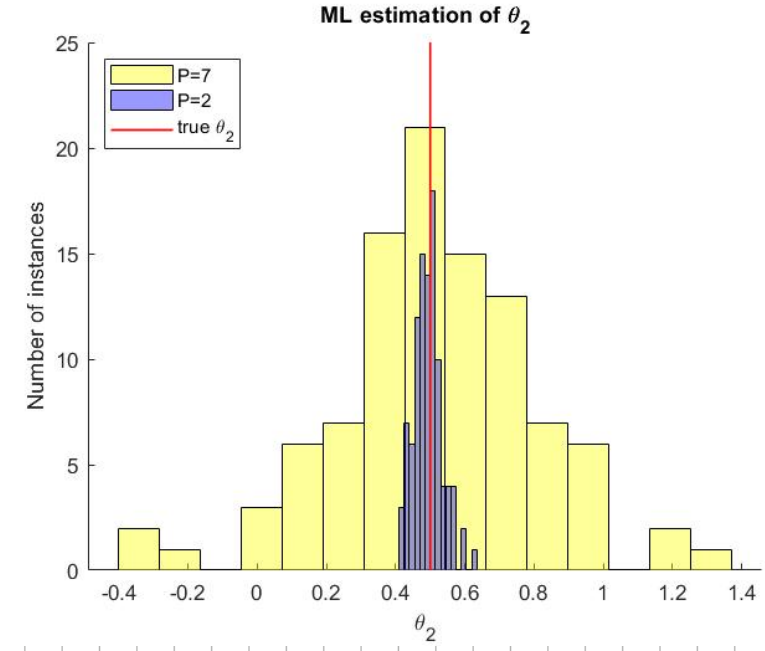
$P=7$: $-1,0 \cdot 10^7$ → enormous since the weights computed in (d) are gigantic.

Here we can see that the maximum likelihood value is achieved at $P=2$, as expected, since $P=1$ underfitted and $P=7$ overfitted, when giving new test data the result explains itself alone.

(8)







As expected, it is possible to see how the the number of parameters is more consistent for the case P=2. We have got the lowest variance for P=2 and the parameters are concentrated around the true value. For P=7 we have got the highest variance in the values of the estimation, this is caused by the fact that we are overfitting and and the weights of the regression get huge (at least for the thetas from 3 to 7, that are not shown here since we have already seen that they are very big, and it does not make sense to plot them since we cannot compare them with P=1 or P=2).

7.2 $\theta = (\theta_0, \dots, \theta_p, \dots, \theta_P)^T$, $p(\theta) = \frac{1}{\sqrt{|2\pi\Sigma_0|}} \exp\left(-\frac{1}{2}(\theta-\mu_0)^T\Sigma_0^{-1}(\theta-\mu_0)\right)$

$\Sigma_0 = I$, $\mu_0 = 0$

(a) $p(\theta|Y) = \frac{p(Y|\theta)\,p(\theta)}{p(Y)}$ ∝ $p(Y|\theta)\,p(\theta) =$

$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \cdot \prod_{i=1}^{N} \exp\left(-\frac{(Y_i - (\theta_0 + \dots + \theta_p x_i^p))^2}{2\sigma^2}\right) \cdot \frac{1}{\sqrt{2\pi}(|\Sigma_0|)^{1/2}} \exp\left(-\frac{1}{2}(\theta-\mu_0)^T\Sigma^{-1}(\theta-\mu_0)\right)$

$\mathcal{L}(p(\theta|Y)) = -N\log(\sqrt{2\pi}\sigma^2) - \sum_{i=1}^{N} \frac{(Y_i - (\theta_0 + \theta_1 x_i + \dots + \theta_p x_i^p))^2}{2\sigma^2} +$

$- \log\sqrt{2\pi} - \frac{1}{2}\log(|\Sigma_0|) - \frac{1}{2}(\theta-\mu_0)^T\Sigma^{-1}(\theta-\mu_0)$

(b) $\frac{d\,\mathcal{L}(p(\theta|Y))}{d\theta} \overset{!}{=} 0$

$\rightarrow \frac{\partial}{\partial\theta}\left(\frac{-1}{2\sigma^2}\|\underline{Y} - \underline{\underline{X}}\,\underline{\theta}\|^2 - \frac{1}{2}(\underline{\theta}-\underline{\mu_0})^T\underline{\underline{\Sigma}}^{-1}(\underline{\theta}-\underline{\mu_0})\right) = 0$

$\rightarrow \frac{\partial}{\partial\theta}\left(\frac{-1}{2\sigma^2}(\underline{Y} - \underline{\underline{X}}\,\underline{\theta})^T(\underline{Y} - \underline{\underline{X}}\,\underline{\theta}) - \frac{1}{2}\theta^T\theta\right) = 0$

$\rightarrow \frac{\partial}{\partial\theta}\left(\frac{-1}{2\sigma^2}\left(Y^TY - \underline{X}^T\underline{\underline{X}}\,\underline{\theta} - \theta^T\underline{\underline{X}}^TY + \theta^TX^TX\theta\right) - \frac{1}{2}\theta^T\theta\right) = 0$

$\frac{1}{2\sigma^2}\left(+2X^TY - 2X^TX\theta\right) - \theta = 0$

$X^TY - X^TX\theta - \theta\sigma^2 = 0 \rightarrow \left(X^TX + \sigma^2\underline{\underline{I}}\right)\theta = X^TY$

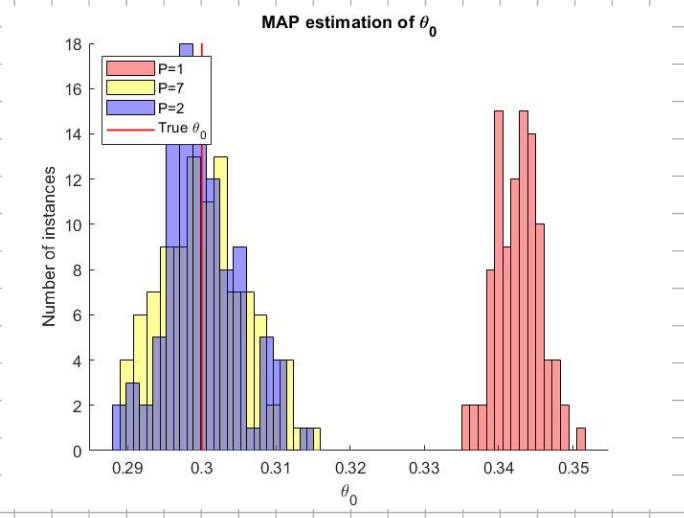
$\rightarrow \boxed{\hat{\theta} = (X^TX + \sigma^2\underline{\underline{I}})^{-1}X^TY}$

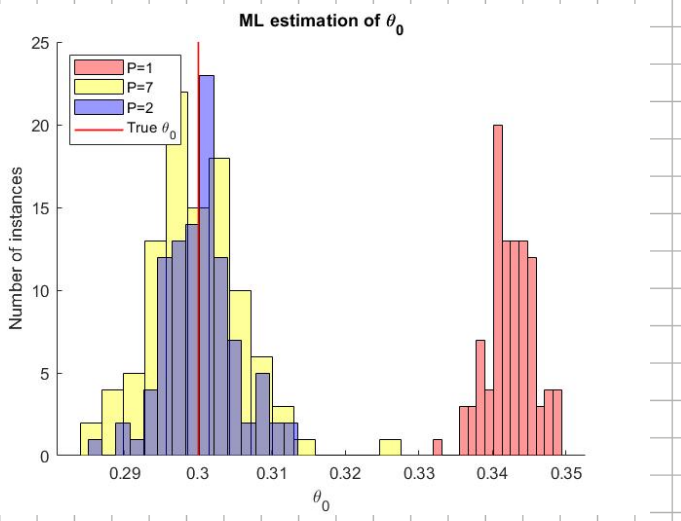
(c) We can see that in the MAP case the variance of the parameters is smaller compared to the ML estimate.

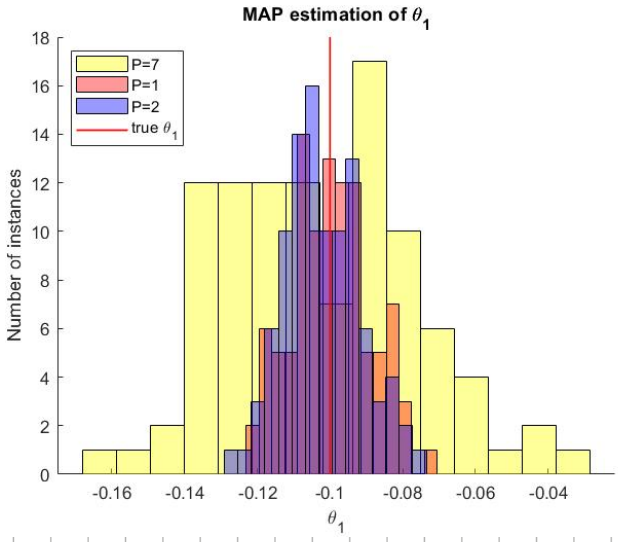
Not only, in fact, if we look at one instance of $\hat{\theta}$ for $p=7$: $\hat{\theta}$ = 0.2963, -0.1160, 0.4756, 0.1615, 0.0673, -0.6167 0.1755, -0.2990

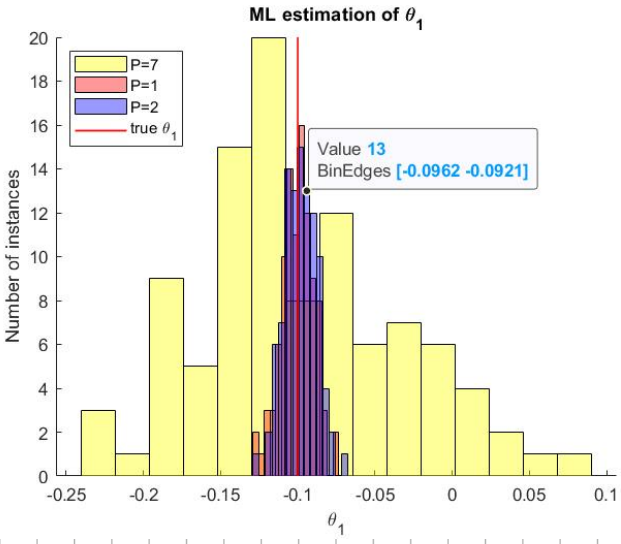
we can see that the parameters from $\theta_3 - \theta_7$ are much smaller now compared to the ML case, this is caused by the regularizing effect of MAP (in this case is ridge regression) on the parameters.

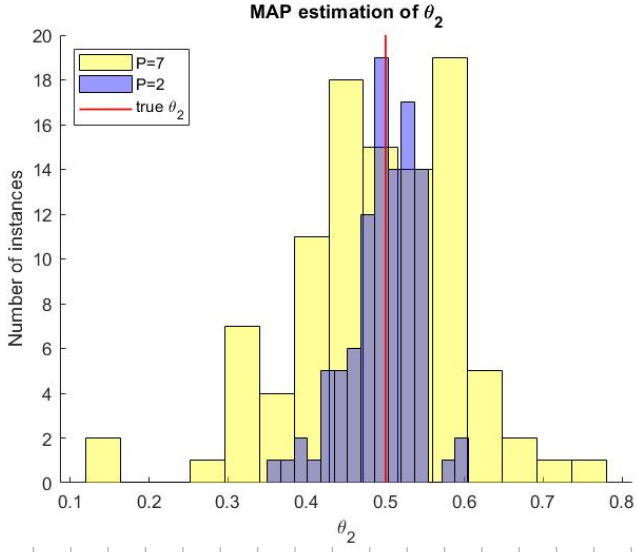
(d)

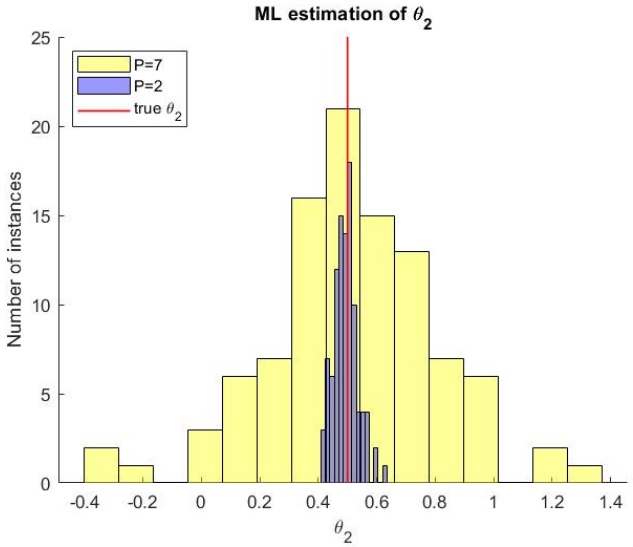












It is noticeable the fact that the parameters for MAP estimation are much more concentrated around the true parameter, the scale of the X-axis is much smaller compared to the one of the ML estimation. This is caused by the sigma^2, that is proportional to the regularization factor (lambda) for ridge regression. By increasing lambda we strengthen the regularization, but in this case also the noise (epsilon) of the dataset.
In conclusion MAP prevents overfitting by keeping the estimated parameters small. (If we set sigma^2 to e.g 100, we can see that the parameters concentrate around 0, get very small, and of course wrong).

1.3 $y = x\theta + \varepsilon$, $p(\varepsilon) \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, $p(\theta) \sim \mathcal{N}(\mu_p, \sigma_p^2)$

(a) $\varepsilon = x\theta - y \rightarrow p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{(x\theta - y)^2}{2\sigma_\varepsilon^2}\right)$

$\mathcal{L}(\theta) = p(y|\theta)$

If we have a Dataset $\{y_i\}_{i=1}^N$

$$\rightarrow \mathcal{L}(\theta) = \left(\frac{1}{2\pi\sigma^2}\right)^N \prod_{i=1}^N \exp\left(-\frac{(x\theta - y_i)^2}{2\sigma_\varepsilon^2}\right) =$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^N \exp\left(-\frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^N x^2\theta^2 - 2x\theta y_i + y_i\right) =$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^N \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\left(Nx^2\theta^2 - 2x\theta N\bar{y} + N\bar{y}\right)\right)$$

(b) $p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{\sqrt{\varepsilon}\,p(y)} \propto p(y|\theta)\,p(\theta)$

$$\rightarrow p(\theta|y) \propto \left(\frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}}\right)^N \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\left(Nx^2\theta^2 - 2x\theta N\bar{y} + N\bar{y}\right)\right) \cdot$$

$$\cdot \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(\frac{(\theta - \mu_p)^2}{2\sigma_p^2}\right) = \frac{1}{(2\pi\sigma_\varepsilon^2)^{N/2}(2\pi\sigma_p^2)^{1/2}} \cdot$$

$$\cdot \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\left(Nx^2\theta^2 - 2x\theta N\bar{y} + N\bar{y}\right) - \frac{\theta^2 - 2\theta\mu_p + \mu_p^2}{2\sigma_p^2}\right)$$

$$\rightarrow \log p(\theta|y) \propto -\frac{N}{2}\log(2\pi\sigma_\varepsilon^2) - \frac{1}{2}\log 2\pi\sigma_p^2 +$$

$$-\frac{1}{2\sigma_\varepsilon^2}\left(Nx^2\theta^2 - 2x\theta N\bar{y} + N\bar{y}\right) - \frac{\theta^2 - 2\theta\mu_p + \mu_p^2}{2\sigma_p^2}.$$

(c) $\log \mathcal{N}(\theta \mid \mu, \sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\theta^2 - 2\mu\theta + \mu^2)$

Take the "exp" part $\propto -\frac{1}{2\sigma_\varepsilon^2\sigma_p^2}\Big(\theta^2(N\bar{x}^2\sigma_p^2 + \sigma_\varepsilon^2) +$

$-2\theta(x N\bar{y}\sigma_p^2 + \mu_p\sigma_\varepsilon^2) + (N\bar{y}^2\sigma_p^2 + \mu_p^2\sigma_\varepsilon^2)\Big)$

The expression in (b) is still a Gaussian, with diffrent μ, σ^2.
Since the conjugate of a Gaussian is a Gaussian, if we multiply them we still get a Gaussian distribution with different mean and variance.
Like the Beta distribution is conjugate to the Binomial distribution.

(d) $-\frac{N}{2}\log(2\pi\sigma_\varepsilon^2) - \frac{1}{2}\log(2\pi\sigma_p^2) - \frac{1}{2\sigma_\varepsilon^2\sigma_p^2}(\theta^2(N\bar{x}^2\sigma_p^2 + \sigma_\varepsilon^2) +$

$-2\theta(x N\bar{y}\sigma_p^2 + \mu_p\sigma_\varepsilon^2) + (N\bar{y}^2\sigma_p^2 + \mu_p^2\sigma_\varepsilon^2))$ (2)

(1) $\frac{-(N\bar{x}^2\sigma_p^2 + \sigma_\varepsilon^2)}{2\sigma_\varepsilon^2\sigma_p^2}\Big(\theta^2 - 2\theta\frac{(x N\bar{y}\sigma_p^2 + \mu_p\sigma_\varepsilon^2)}{N\bar{x}^2\sigma_p^2 + \sigma_\varepsilon^2} + \dots$

. . . . Sorry, too long.